BMJ

BMJ 2012;345:e5840 doi: 10.1136/bmj.e5840 (Published 14 September 2012)

RESEARCH

Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis

 OPEN ACCESS

Brennan C Kahan *medical statistician*, Tim P Morris *medical statistician*

MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London WC2B 6NH, UK

Abstract

Objectives To assess how often stratified randomisation is used, whether analysis adjusted for all balancing variables, and whether the method of randomisation was adequately reported, and to reanalyse a previously reported trial to assess the impact of ignoring balancing factors in the analysis.

Design Review of published trials and reanalysis of a previously reported trial.

Setting Four leading general medical journals (*BMJ*, *Journal of the American Medical Association*, *Lancet*, and *New England Journal of Medicine*) and the second Multicenter Intrapleural Sepsis Trial (MIST2).

Participants 258 trials published in 2010 in the four journals. Cluster randomised, crossover, non-randomised, single arm, and phase I or II trials were excluded, as were trials reporting secondary analyses, interim analyses, or results that had been previously published in 2010.

Main outcome measures Whether the method of randomisation was adequately reported, how often balanced randomisation was used, and whether balancing factors were adjusted for in the analysis.

Results Reanalysis of MIST2 showed that an unadjusted analysis led to larger P values and a loss of power. The review of published trials showed that balanced randomisation was common, with 163 trials (63%) using at least one balancing variable. The most common methods of balancing were stratified permuted blocks (n=85) and minimisation (n=27). The method of randomisation was unclear in 37% of trials. Most trials that balanced on centre or prognostic factors were not adequately analysed; only 26% of trials adjusted for all balancing factors in their primary analysis. Trials that did not adjust for balancing factors in their analysis were less likely to show a statistically significant result (unadjusted 57% *v* adjusted 78%, P=0.02).

Conclusion Balancing on centre or prognostic factors is common in trials but often poorly described, and the implications of balancing are poorly understood. Trialists should adjust their primary analysis for balancing factors to obtain correct P values and confidence intervals and to avoid an unnecessary loss in power.

Introduction

The randomised controlled trial is considered the ideal study design for assessing the effect of an intervention, as it is the only method of ensuring that no systematic differences exist between treatment groups. However, differences between treatment arms in important prognostic factors can still arise by chance. Such differences may cause some to question the validity of the trial results. Many trials use balanced randomisation to ensure a similar distribution between treatment groups in important variables thought to influence outcome, such as age and disease stage. Balanced randomisation involves selecting certain baseline covariates (called balancing variables) and incorporating them into the randomisation scheme in a way that forces a certain degree of balance between treatment arms. Common methods of balancing are minimisation[1] or permuted blocks within strata.[2]

There is overwhelming evidence from the statistical literature to show that variables used in the randomisation process should subsequently be adjusted for in the analysis,[3-8] as failure to do so can result in P values that are too large and confidence intervals that are too wide; this leads to a decrease in power and a reduction in type I error rate, which could potentially lead to an incorrect conclusion that the treatment has no benefit. This is because balanced randomisation introduces correlation between treatment groups, which violates the statistical assumption that all patients are independent.[4] This correlation between treatment groups occurs because balanced randomisation forces the outcomes between treatment arms to be similar (apart from any treatment effect). This is seen in figure 1⇓, where simulated data shows that outcomes between the two treatment groups are correlated under balanced

Correspondence to: B C Kahan brk@ctu.mrc.ac.uk

Extra material supplied by the author (see http://www.bmj.com/content/345/bmj.e5840?tab=related#webextra)

Differences between adjusted and unadjusted analyses
Differences in sample sizes by randomisation method

randomisation. A more detailed discussion of these issues can be found elsewhere.[4]

It is therefore important to adjust for all balancing factors in the analysis to ensure correct P values and confidence intervals. Giving a clear description of both the method of randomisation and the method of analysis is also important to allow readers to properly judge trial results.[9] For example, if the method of randomisation is unclear, readers will be unable to judge whether the method of analysis was adequate. We assessed the potential impact of unadjusted analyses after balanced randomisation had been used; determined how often trials balance on prognostic factors or recruiting centre, and whether these factors were appropriately adjusted for in the analysis. We also determined whether the method of randomisation was adequately reported.

Methods

Reanalysis of the MIST2 trial

To assess the impact of an unadjusted analysis after balanced randomisation has been used, we reanalysed data from a published randomised trial. The second Multicenter Intrapleural Sepsis Trial (MIST2)[10] found that a combination of tissue plasminogen activator and DNase was effective in reducing the size of pleural effusions in patients with pleural infection. Patients were randomised using minimisation, balancing on the size of the pleural effusion at baseline, whether the infection was acquired in hospital or not, and the presence of purulent pleural fluid.

We reanalysed the primary outcome (size of pleural effusion) and the two major secondary outcomes (need for surgery and time to hospital discharge). We used both adjusted and unadjusted analyses and compared the results.

Review of trials published in leading medical journals

One author (BCK) searched the electronic table of contents of the *BMJ*, *Journal of the American Medical Association*, *Lancet*, and *New England Journal of Medicine* between January and December 2010 for reports of parallel group, individually randomised trials. We discarded articles with titles that indicated non-randomised trials. All other articles were downloaded and assessed for eligibility. Cluster randomised, crossover, single arm, and phase I or II trials were excluded, as were non-randomised studies. We excluded cluster randomised trials because of possible differences compared to individually randomised trials in types of balancing variables used, and we excluded phase I and II trials as we wanted to focus on large scale phase III trials that had the ability to change clinical practice. To avoid double counting we additionally excluded articles reporting secondary analyses, interim analyses, or results that had been previously published in 2010. A second author (TPM) repeated this process for articles published between March and May 2010 to assess agreement in the articles identified and those classified as eligible. Agreement between authors was 100% for both.

We extracted data onto a standardised form, which was piloted on several articles from 2009. One author (BCK) extracted data from all trials and the other author (TPM) extracted data from 20 randomly selected trials to assess agreement between authors. Discrepancies were resolved by discussion. Agreement was assessed for whether the allocation ratio was specified, the method of randomisation, whether centre was used as a balancing factor and adjusted for in the analysis, and whether any prognostic variables were used as balancing factors and adjusted for in the analysis. Overall agreement between the two authors was 96%.

We chose to review articles from the four selected medical journals to enable comparisons with previously reported reviews of the same journals.[11 12] We reviewed articles published during one calendar year because a previous report[12] found over 200 eligible trials during the same time frame from the same group of journals, which would allow good estimates of how often balanced randomisation is used and how often the balancing factors are accounted for in the analysis.

Trials were classified as adjusting for balancing factors if the article stated that balancing factors had been adjusted for or if they included the balancing factors in a list of adjustment factors. The trials were classified as not adjusting if they stated an unadjusted analysis was done or if they listed those factors that had been adjusted for but did not list the balancing factors. For trials that provided summary outcome information (means and standard deviations or proportions in each group) we calculated the crude (unadjusted) treatment effect. If this result matched the treatment effect given in the text, we listed the trial as unadjusted. Otherwise the trial was listed as unclear. We listed as unclear any additional trials that were not classified as adjusted or unadjusted.

If both adjusted and unadjusted results were presented but neither was classified as the primary analysis, we took whichever result was presented first as the primary analysis. When the primary outcome was not stated, we took it to be the first outcome listed.

Results

Reanalysis of MIST2

Supplementary table 1 shows the differences between adjusted and unadjusted analyses. Unadjusted analyses led to larger P values for need for surgery (a 1.8-fold increase; 0.095 *v* 0.175 for adjusted and unadjusted, respectively) and time to hospital discharge (a fourfold increase; 0.011 *v* 0.044). For the size of the pleural effusion both analysis methods gave identical P values (P=0.005); however, this was because the unadjusted analysis had a larger treatment effect, owing to a baseline imbalance in the size of the pleural effusion between treatment groups. The width of the confidence interval for the unadjusted analysis, however, was 56% wider than that of the adjusted analysis, which could lead to a reduction in power of over 20%.[4]

The unadjusted analysis for need for surgery led to a smaller confidence interval than the adjusted analysis. This phenomenon has been explained previously[13]; for binary and time to event outcomes, adjusting for prognostic factors increases both the standard error and the estimated treatment effect (given there is a treatment effect). This leads to wider confidence intervals but smaller P values, meaning that adjusted analyses will still increase power for binary and time to event outcomes. Previous studies have shown that if balancing or prognostic factors are well chosen, increases in power can be substantial (for example, >10%).[4 13 14] Although the confidence interval for the unadjusted analysis is smaller than that of the adjusted analysis, it is still incorrect in the sense that the type I error rate will be smaller than it should be; by comparison the type I error rate for the adjusted analysis will be correct.

Review of trials published in leading medical journals

Overall, 304 trials were identified, of which 46 were excluded: 17 were cluster randomised trials, 11 phase I or II trials, nine

previously reported trials from the same year, four were crossover trials, three carried out secondary analyses, one did an interim analysis, and one was a non-randomised trial. In total 258 trials were included (fig 2⇓). Table 1⇓ shows the characteristics of the included trials.

Randomisation

In 96 trials (37%) the method of randomisation was unclear. Among those trials that reported the method of randomisation, four (2%) used simple randomisation, 125 (77%) used permuted blocks (85 with stratification and 40 without), 29 (18%) used minimisation, and four (2%) used another method.

Among trials using permuted blocks, 42 (34%) did not state the block size. Fifteen trials (12%) used random block sizes, whereas in 42 trials (34%) this was not clear. The median block size used (taking the largest when random block sizes were used) was 8 (interquartile range 4-10, 10-90th centile 4-20). Sixty nine trials (83%) used a block size less than 12.

Twenty seven of 29 trials (93%) that used minimisation did not specify whether it was deterministic—that is, completely non-random. The two trials that reported using an element of probability did not state the probability of receiving the favoured treatment.

Use of balanced randomisation

Randomisation was balanced on centre in 120 trials (47%) and on prognostic factors in 111 trials (43%). In total, 163 trials (63%) balanced on at least one variable (centre or a prognostic factor). Most trials balanced on only one or two prognostic factors (n=87; 78%), whereas 24 trials (22%) used between three and eight factors.

Analysis of trials using balanced randomisation

Of those trials that balanced on centre, only 31 (26%) reported adjusting the primary analysis for centre; 4 (3%) adjusted for centre in a secondary analysis, 68 (57%) did not adjust for centre, and 17 (14%) were unclear (table 2⇓).

Similarly, only 40 trials (36%) that used prognostic factors in their randomisation adjusted for all of these factors in the primary analysis; 4 (4%) adjusted for some factors, 10 (9%) adjusted for all factors in a secondary analysis, 45 (41%) did not adjust for any factors, and 12 (11%) were unclear.

Overall, only 42 (26%) of trials that used at least one balancing factor in their randomisation (either centre or a prognostic factor) appropriately adjusted for all factors in the primary analysis; 8 (5%) adjusted for all factors in a secondary analysis, 3 (2%) adjusted for centre but not for prognostic factors, 14 (9%) adjusted for prognostic factors but not centre, 74 (45%) did not adjust for any balancing factors, and 22 (14%) were unclear.

Three of 10 trials that adjusted for all prognostic factors in a secondary analysis, two of four trials that adjusted for centre in a secondary analysis, and three of eight trials that adjusted for all balancing factors in a secondary analysis gave equal weight to both the adjusted and unadjusted analyses, but presented the unadjusted analysis first. The remaining trials gave more weight to the unadjusted analysis—for example, by presenting only unadjusted results in the abstract, text, or key figures or tables. None of these trials had specified which of the two analysis methods was the primary.

Difference in significance rates between adjusted and unadjusted analyses

Trials that adjusted for all balancing factors were more likely to find a statistically significant result (n=39/50, 78%) compared with trials that did not adjust for any balancing factors (n=42/74, 57%; odds ratio (adjusted *v* unadjusted) 2.70, P=0.02), indicating that trials could be losing power through not adjusting. These results should be interpreted cautiously, however, as this difference could in part be due to confounding. Fourteen of 22 (64%) trials that did not clarify whether the analysis was adjusted found a statistically significant result, as did 12/17 (71%) trials that adjusted for some but not all balancing factors.

Discussion

Balanced randomisation introduces correlation between treatment groups, violating the statistical assumption that all observations are independent. Accounting for balancing factors in the analysis is necessary to obtain correct P values and avoid a loss in power. Most trials reviewed (63%) used balanced randomisation; however only 26% appropriately adjusted for all balancing factors in their primary analysis, indicating that the majority of trials using balanced randomisation may be reporting overly conservative results.

Reporting of randomisation

Many trial reports made no attempt to explain the method of randomisation and simply reported that patients were randomised to different treatments (see box for examples). Other trials attempted to explain the method of randomisation but did so poorly, stating only that randomisation was done using computer generated random numbers or a random numbers table. These explanations are not adequate as almost all methods of randomisation can be done using random numbers. The method of randomisation can have a large effect on the possibility of selection bias in open label trials,[12 15 16] and the appropriate method of analysis also depends on the method of randomisation. Therefore the method of randomisation should be clearly explained so that readers are able to determine whether appropriate methods were used. One policy that would add clarity to reporting is for authors to explicitly state when randomisation was not stratified (see seventh example in box) and when an unadjusted analysis was done. Currently, most trials only mention stratification when it was used but do not explicitly state when it was not used. This makes it difficult to judge whether stratification was truly not used or if it was used but was not reported.

Analysis of trials after stratified randomisation

We have found that ignoring balancing factors in the analysis after stratified randomisation can impact on trial results. In the second Multicenter Intrapleural Sepsis Trial (MIST2), ignoring the balancing factors led to a 56% increase in the width of the confidence interval for the primary outcome (which could lead to a reduction in power of over 20%) and led to 1.8-fold and 4.0-fold increases in the P values for need for surgery and time to discharge, respectively. These results are consistent with an example given by a randomised trial comparing two chemotherapy treatments for liver cancer.[6] Randomisation was balanced across 18 centres; ignoring centre in the analysis led to a 4.5 fold increase in the P value (0.027 *v* 0.006 for unadjusted and adjusted, respectively).

We found that 71% of trials that balanced on centre and 55% that balanced on prognostic factors did not appropriately adjust

Explanations of randomisation method

Poor explanations of randomisation method

Patients were then randomly assigned to [treatment] or placebo for 12 weeks

Patients were randomly assigned to receive usual care or [treatment], according to a sequence of computer-generated random numbers, with stratification on the basis of the study site

Qualifying participants underwent randomisation and started the assigned study medication as inpatients

Within two weeks of recruitment we randomly allocated sealed sample packs . . . into two groups using random number tables

Patients were randomly assigned by a computer program to receive either [treatment] three times a day or [placebo] thrice daily

Good explanations of randomisation method

Eligible subjects were randomly assigned to receive one of the three study medications in a 1:1:1 ratio. Treatment assignments were performed centrally according to a computer-generated random schedule in permuted blocks of three within age strata (<6 years and ≥6 years) and within study site

After providing written informed consent, [patients] were randomly assigned in a 1:1:1 ratio (with the use of sealed envelopes) to one of three study groups in permuted blocks of six or nine with no stratification

Patients were randomly assigned (1:1) by simple randomisation to the [intervention] or to standard care (control group). A project statistician generated the randomisation numbers with a random number generating program

for all factors in their analyses. Trials that adjusted for balancing factors in the analysis were more likely to show a statistically significant result, potentially because of the increased power owing to adjustment.

Comparison with other studies

Two previous reviews in 1997 and 2002 analysed the same journals we reviewed and found the method of randomisation was unclear in 54% and 34% of trials, respectively.[11 12] The findings from one of the reviews are similar to our own (37% of trials did not specify the method of randomisation), indicating that while reporting of the randomisation method may have improved between 1997 and 2002 there has been little improvement subsequently, which we hope will change with the adoption of the 2010 consolidated standards of reporting trials (CONSORT) statement.[9 12] We found the number of trials balancing for centre and prognostic factors to be similar to that reported previously.[11]

Limitations of the study

Our review was limited to articles published in four major medical journals, which is unlikely to be a representative sample as articles published in other medical journals are likely to have different reporting standards. Journals adopting the CONSORT statement have better reporting standards than others,[17] so it is likely that overall reporting standards are worse than those we found. The majority of trials were identified and reviewed by one author, with only a subset of trials identified or reviewed by a second author. Although agreement between authors was high (100% for trial identification, 96% for data extraction), human error remains possible.

Conclusions and policy implications

Balanced randomisation induces correlation between treatment groups, which can lead to P values that are too large and confidence intervals that are too wide if balancing factors are not accounted for in the analysis. This is unlikely to affect trials that show overwhelming evidence in favour of a treatment difference, but could affect interpretations in the presence of moderate evidence. The difference between adjusted and unadjusted analyses depends on whether the balancing factors are associated with outcome. If balancing factors are not prognostic, then ignoring them in the analysis will have little impact; however, balancing factors are generally chosen because they are thought to be prognostic and so should generally be associated with outcome. Because the analysis method should be prespecified before data analysis (as retrospective model selection where authors use statistical significance tests to determine which factors should be adjusted for can lead to type I error rates that are too high[7]), we recommend the protocol or statistical analysis plan should prespecify that all balancing factors are adjusted for in the analysis.

We thank Daniel Bratton, Ben Sear, and Najib Rahman for their comments on the manuscript; the MIST2 trial team for the use of their data; and the journal editors and reviewers whose comments improved the manuscript.

Contributors: BCK devised the study, designed the data extraction forms, extracted data from all trial reports, tabulated the results, and wrote the first draft of the manuscript. He is guarantor. TPM contributed to the design of the data extraction forms, extracted data from 20 trial reports, and contributed to the writing of the manuscript.

Funding: Both authors are employed by the MRC Clinical Trials Unit. TPM is funded by an MRC studentship (MC-US-A737-0012).

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: A full list of the trials reviewed and data abstracted for each trial is available from the corresponding author at brk@ctu.mrc.ac.uk.

1 Altman DG, Bland JM. Treatment allocation by minimisation. *BMJ* 2005;330:843.
2 Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. *Control Clin Trials* 1988;9:327-44.
3 ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. *Stat Med* 1999;18:1905-42.
4 Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med* 2012;31:328-40.
5 Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol* 1999;52:19-26.
6 Parzen M, Lipsitz SR, Dear KBG. Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial?. *Biom J* 1998;40:385-402.
7 Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. *Control Clin Trials* 2000;21:330-42.
8 Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials. a review. *Control Clin Trials* 2002;23:662-74.
9 Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
10 Rahman NM, Maskell NA, West A, Teoh R, Arnold A, Mackinlay C, et al. Intrapleural use of tissue plasminogen activator and DNase in pleural infection. *N Engl J Med* 2011;365:518-26.
11 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064-9.
12 Hewitt CE, Torgerson DJ. Is restricted randomisation necessary? *BMJ* 2006;332:1506-8.

What is already known on this topic

Stratified randomisation introduces correlation between treatment groups

Ignoring stratification factors in the analysis leads to P values that are too large and to incorrect confidence intervals

Adjusted analyses increase power and give correct type I error rates

What this study adds

Not adjusting for the stratification factors in the second Multicenter Intrapleural Sepsis Trial led to larger P values and wider confidence intervals for some outcomes

Stratified randomisation is common, with 63% of trials reporting the use of at least one stratification factor

Only 31% of trials using stratified randomisation appropriately adjusted for all factors in their analysis

13 Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004;57:454-60.
14 Turner EL, Perel P, Clayton T, Edwards P, Hernandez AV, Roberts I, et al. Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury. *J Clin Epidemiol* 2012;65:474-81.
15 Kennes LN, Cramer E, Hilgers RD, Heussen N. The impact of selection bias on test decisions in randomized clinical trials. *Stat Med* 2011;30:2573-81.
16 Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. *Lancet* 2002;359:966-70.
17 Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust* 2006;185:263-7.

Accepted: 20 August 2012

Cite this as: *BMJ* 2012;345:e5840

Tables

Table 1| Characteristics of included trials. Values are number (percentage) of trials unless stated otherwise

Characteristics	Trials (n=258)
Median (interquartile range) No of patients	557 (232-1679)
No of treatment arms:	
2	213 (83)
3	34 (13)
≥4	11 (5)
Primary outcome:	
Continuous	64 (25)
Binary	100 (39)
Time to event	81 (31)
Rate	11 (4)
Other	2 (1)
No of centres involved:	
Single	22 (9)
Multiple	206 (80)
Not stated	30 (12)
Allocation ratio:	
1:1	110 (43)
Not 1:1	19 (7)
Not stated*	129 (50)

*Most trials that did not state the allocation ratio had a similar number of patients in all treatment arms, indicating the ratio was likely to be 1:1

Table 2| **Reporting and analysis of trials**

Variables	No (%) of trials (n=258)
Method of randomisation:	
Simple randomisation	4 (2)
Permuted blocks without stratification	40 (16)
Permuted blocks with stratification	85 (33)
Minimisation	29 (11)
Other	4 (2)
Unclear	96 (37)
Balanced on centre	120 (47)
Adjustment for centre (n=120)*:	
Adjusted primary analysis	31 (26)
Adjusted secondary analysis	4 (3)
Not adjusted	68 (57)
Unclear	17 (14)
Balanced on prognostic factors	111 (43)
Adjustment for prognostic factors (n=111)*:	
Adjusted primary analysis for all factors	40 (36)
Adjusted primary analysis for some factors	4 (4)
Adjusted secondary analysis for all factors	10 (9)
Not adjusted	45 (41)
Unclear	12 (11)
Balanced on centre or prognostic factors	163 (63)
Adjustment for centre or prognostic factors (n=163)*:	
Adjusted primary analysis for all factors	42 (26)
Adjusted secondary analysis for all factors	8 (5)
Adjusted for centre but not prognostic factors	3 (2)
Adjusted for prognostic factors but not centre	14 (9)
Not adjusted	74 (45)
Unclear	22 (14)
No of prognostic factors balanced on:	
1	53 (48)
2	34 (31)
3	11 (10)
4	8 (7)
5	2 (2)
8	3 (3)

*Trials were only assessed for whether they adjusted for centre or prognostic factors if they had balanced on these factors.

Figures

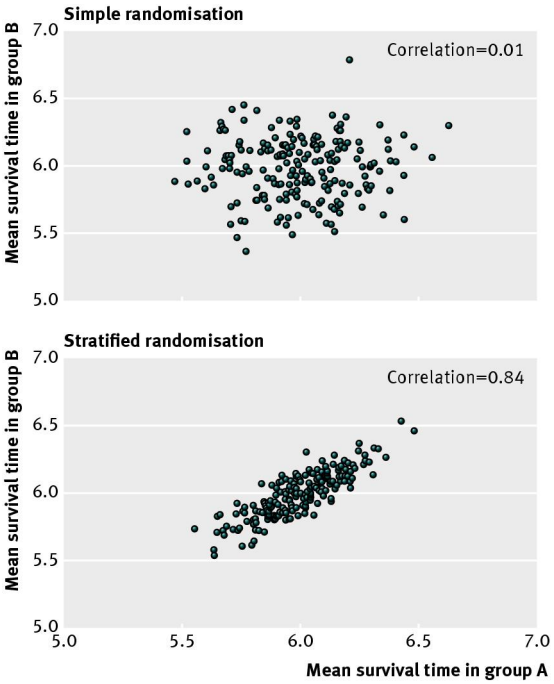

Fig 1 Correlation in mean survival time between treatment groups under simple and stratified randomisation (simulated data). Data were generated from the formula: survival time=3months+(6months)×(early disease stage)+random error, where random error ~ N(0, 1)

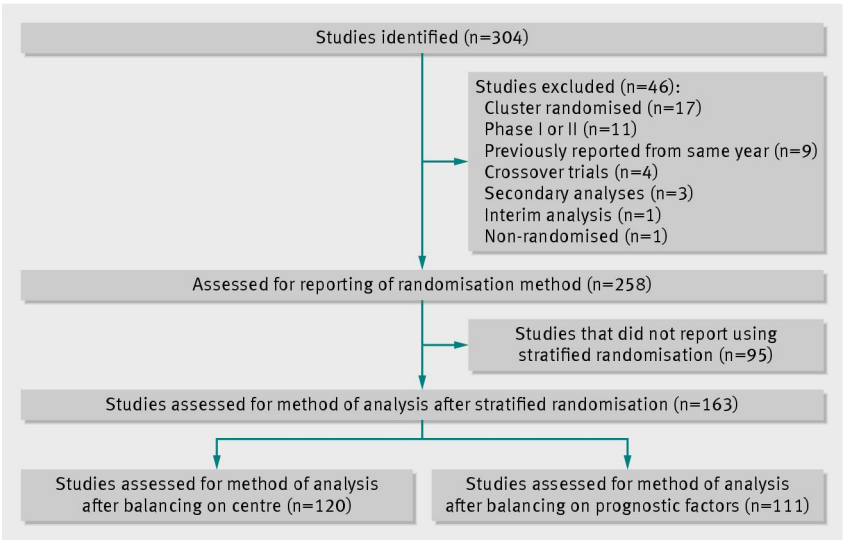

Fig 2 Flow diagram of study selection